Salient Montages from Unconstrained Videos

Min Sun, Ali Farhadi, Ben Taskar, and Steve Seitz

University of Washington

Fig. 1: Given a video (click to watch on Youtube: link) captured by a head-mounted camera (top row), we first automatically identify montageable moments (highlighted by the color-coded bounding boxes) containing the salient person (the little girl in pink) and ignore irrelevant frames. A set of salient montages ordered by our novel montageability scores is generated automatically. Here we show four typical examples.

Abstract. We present a novel method to generate salient montages from unconstrained videos, by finding "montageable moments" and identifying the salient people and actions to depict in each montage. Our method addresses the need for generating concise visualizations from the increasingly large number of videos being captured from portable devices. Our main contributions are (1) the process of finding salient people and moments to form a montage, and (2) the application of this method to videos taken "in the wild" where the camera moves freely. As such, we demonstrate results on head-mounted cameras, where the camera moves constantly, as well as on videos downloaded from YouTube. Our approach can operate on videos of any length; some will contain many montageable moments, while others may have none. We demonstrate that a novel "montageability" score can be used to retrieve results with relatively high precision which allows us to present high quality montages to users.

Keywords: video summarization, video saliency detection

1 Introduction

Video is increasingly easy to capture and store. The advent of wearable devices like GoPro cameras is accelerating this trend, allowing us to capture hours at a time in a hands-free manner. While the vast majority of this footage is unlikely to be interesting or useful, the hope is that if something interesting *does* happen, we will have recorded it, and be able to generate an *at-a-glance* visualization. Finding those interesting moments, however, is like looking for a needle in a haystack, and motivates the need for video search and summarization research.

Finding semantically interesting moments via automated means is extremely challenging. Instead, we seek to find moments that *look interesting*, and, in particular, produce high quality photo montages fully automatically (see Fig. 1).

Each montage captures a *stroboscopic* image of a person performing an action, with the same person shown multiple times in the same image as if a strobe light had flashed multiple times during the same exposure. Pioneered in the 19th century by Etienne-Jules Marey, stroboscopic images provide a fascinating "time-lapse" view of an action in a single image. While Marey's work required a special chronophotographic gun, modern solutions [1,31,17,41] enable similar results with regular photos via algorithmic means. These techniques require as input several photos or a short video clip comprising the *montageable* event. The camera must remain still or only pan slowly in order to create effective montages.

Nevertheless, most videos do not satisfy these constraints; Hence, automatically producing montages from unconstrained videos is extremely challenging. For example, such videos often contain crowded scenes with many people (see Fig. 4(b)). Existing methods lack the information to select the salient person to depict. Moreover, when the camera moves freely and/or the depth variation of the scene is large (see the playground scene in Fig. 1), global registration methods will fail; hence, low-level motion cues become unreliable.

In this work, we propose a novel, human-centric method to produce montages from unconstrained videos, by finding "montageable moments" and identifying salient people and actions to depict in each montage. Our contribution is not the compositing algorithm itself, which builds upon [1], but (1) the process of finding salient people and moments to form a montage, and (2) the application of this method to videos "in the wild." As such, we demonstrate results on videos from head-mounted cameras, where the camera moves constantly, as well as on videos downloaded from YouTube. The videos from head-mounted camers are particularly challenging since they are unedited and include many irrelevant moments due to motion blur from fast camera movement, self-occlusion from the wearer, and a lot of moments when the wearer is simply navigating the terrain (see Fig. 3). Our approach overcomes all these challenges and can operate on videos many minutes or hours long; some will contain many montageable moments, while others may have none. For this application, we specifically aim to achieve high precision (i.e., a small number of "great" summaries rather than summarizing every moment) from a large number of user videos. Note that high precision is important in many problems such as recommender systems [7].

Our approach is based on (1) clustering people tracklets into "places" (see color-coded boxes in Fig. 1-Top), (2) identifying the most salient people in each place (the little girl in Fig. 1-A), and (3) evaluating the montageability of the people tracklets in each place to select a few hypotheses (four selected person instances in Fig. 1-A). Two key enablers are a new poselet-based human detection and tracking method, and a novel tracklet-based saliency detector. The latter is based on random forests trained on gaze tracking data collected from other videos. We show that this tracklet saliency method outperforms prior saliency techniques for this task. For the third step, we minimize a montageability function that considers the scene complexity and registration quality for all human hypotheses in the place. Finally, we use a novel "montageability" score to rank the quality of the montages, which allows us to present high quality composites to users (see montages in Fig. 1-Bottom).

2 Related Work

In this section, we review related work in video summarization, video saliency detection, and person tracking.

Video summarization: There is a large literature on video summarization (see review [4]), including techniques for sampling important frames [28,22,25,15,33,21] or generating montages. In the following, we discuss the ones most relevant to our method. Aner and Kender [2] automatically generate montages by taking a background reference frame and projecting foreground regions into it. Liu et al. [26] automatically extract panoramas from Youtube videos. However, they assume a panning camera and focus on short clips with few objects. [36,37] focus on extracting highlights from webcams or surveillance cameras and generate synopses which show several spatially non-overlapping actions from different times of the video. However, they assume the camera is stationary so that low-level motion cues can be reliably used to search for salient regions in time and space. Several methods [14,18] involving user interaction have also been proposed in the graphics and HCI communities. All of these methods focus on short clips and assume the camera is mostly stationary. To the best of our knowledge, our method is the first to handle any video, even these captured by a wearable camera.

Saliency detection: Many methods have been proposed to detect salient regions from video. However, most methods [8,16,30,39,38] rely on low-level appearance and motion cues as inputs. A few methods [19,13,38] include information about face, people, or context. Among them, [38] is the state-of-the-art video saliency detector, since it explicitly models the conditional saliency between consecutive frames. However, they have focused primarily on TV series that typically do not contain many people. Unlike our method, they only keep a few candidate regions per frame and do not explicitly solve the person association problem (tracking).

Tracking humans: Many tracking systems are based on linking candidate human hypotheses [35,43,32]. However, these systems obtain inferior performance due to severe body part articulation and camera motion in unconstrained videos. Other works address these issues [12,6] relying on supervoxel and/or long-term point trajectories which are computational expensive to obtain. Our system tracks pre-defined poselets to increase the accuracy without applying additional process (e.g., supervoxel).

3 Our Approach

Videos captured by casual users with smartphones or head mounted cameras are challenging because they typically contain significant camera motion and shake, inconsistent framing and composition, and a lot of redundant content. We posit that the interesting moments in such videos typically involve people, and we focus on extracting person-centric montages that capture their salient actions. We address this goal by identifying salient people in an unconstrained video (Fig. 2-Top) and then generating montages composed of their salient actions (Fig. 2-Bottom). The overview of our approach is depicted in Fig. 2. We first describe how to identify salient people in an unconstrained video.

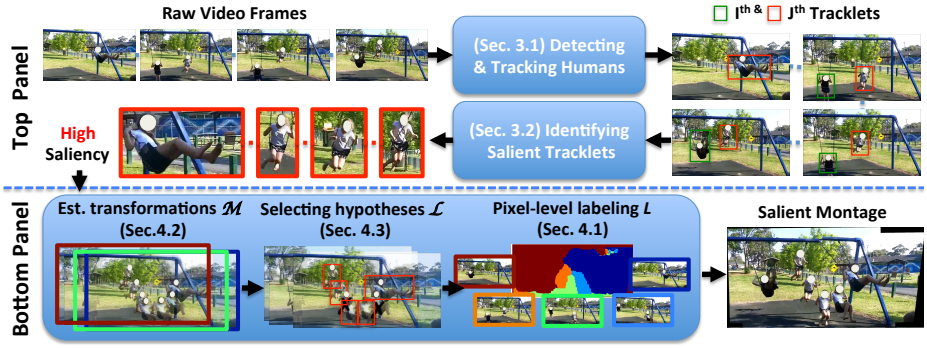


Fig. 2: Detailed system overview (click to watch raw video on Youtube: link). Top panel shows the steps toward identifying salient tracklets (Sec. 3.1 and 3.2), where each tracklet consists of color coded hypotheses. Bottom panel shows the steps toward generating a salient montage (Sec. 4). The color in the pixel-level labeling L indicates the source images indexed by color as well.

Our selection process begins with detecting and tracking all the people in the video, followed by choosing a few salient ones using motion, pose, composition and other cues. Although human detection and tracking are well-studied, our videos pose unique challenges to the state-of-the-art methods, which we address below.

3.1 Detecting and tracking humans

Detecting and tracking humans is critical for our application, since low-level motion cues are unreliable for segmenting out foreground objects due to severe camera motions in unconstrained videos. One of the primary challenges for human detection is high variation in pose and occlusion patterns in unconstrained videos. We found the poselet detector [5] to be robust and particularly useful for our approach. The poselet detector provides human hypotheses (a bounding box represents the extent of a whole human body) along with poselet activations (a poselet is a group of body parts such as left-arm, lower body, etc.), which we use to make tracking more precise. For simplicity, the human hypothesis and poselet activation are referred to as hypothesis and activation, respectively.

Detection by poselet trajectories. For each frame, we begin by matching poselet templates to HOG features [9] extracted at multiple scales and locations to compute poselet activations. However, we do not directly convert these activations into human hypotheses. We track the activations across nearby frames to form poselet trajectories consisting of more reliable activations. Instead of tracking poselet bounding-boxes, which include both background and foreground regions, we track the foreground region of each poselet[1] using a median flow tracker [20]. We start from the first frame by using a set of activations that are sufficiently trackable in the next frame. We repeat the process and form poselet trajectories as detailed in the technical report [40]. At the end of this stage we have a set of activations (original + tracked) in our poselet trajectories, which we

[1] Foreground mask of each poselet is included in the poselet detector.

spatially group into human hypotheses using a Hough voting scheme [3] similar to [5]. We have shown in Table 1(a) that this process significantly increases the detection accuracy.

Tracking by poselet association. Given hypotheses with their poselet activations at each frame, we form tracklets (i.e., a set of hypotheses of the same individual in different frames) by associating hypotheses at consecutive frames. Standard tracking by detection approaches associate hypotheses across neighboring frames by using the location and appearance information in a coarse bounding box representing the extent of a human. We, however, proceed to associate hypotheses using their poselet activation as descriptors. Note that robust hypotheses association is crucial for avoiding identity switch in a tracklet or tracklets drifting to background regions.

Poselet-based similarity. We divide each poselet into 4 by N square cells[2], where 8×3 L_1 normalized color-histogram in Lab space are extracted from each cell. For each hypothesis, we concatenate the poselet histogram following a predefined order to generate hypothesis histogram a. The poselet-based similarity of a pair of hypotheses i and j in two consecutive frames $(t, t+1)$ is defined using the cosine similarity $sim_{ij} = \frac{a_i^T a_j}{\|a_i\| \|a_j\|}$. Although the similarity helps us avoid associating hypotheses with dissimilar poselet activations, it is insufficient to avoid associating false hypotheses fired at non-distinctive background regions. We address this problem by defining a more robust "relative similarity".

Relative similarity. We utilize the smoothness prior on hypotheses locations within each tracklet (i.e., the locations of hypotheses at consecutive frames should be close) to define *relative similarity*. For each hypothesis i at frame t, a subset of hypotheses C at $t+1$ are selected as the candidate set satisfying the smoothness prior if every hypothesis in C at least has ρ spatial overlap with hypothesis i. We define the relative similarity γ_{ij} of a pair of hypotheses i and j as the ratio between sim_{ij} and $\max_{j' \in C'} sim_{ij'}$, where C' is the complement of C. Note that $\max_{j' \in C'} sim_{ij'}$ will be high for a hypothesis i fired on a non-distinctive background region. As a result, false hypotheses fired at non-distinctive background regions tend to have small relative similarity.

Given the relative similarity between candidate pairs of hypotheses, we formulate the hypotheses association problem as a network flow problem and solve it approximated using a dynamic programming algorithm [35]. For simplicity, the costs of the links in the network are the same if the relative similarity γ_{ij} is greater or equal to a threshold σ; otherwise, the costs are infinite. Finally, the network flow problem is efficiently solved to generate tracklets $\{T_k\}_k$ indexed by the tracklet index k. Each tracklet $T_k = \{j\}$ consists of a set of human hypotheses, where j is the hypothesis index. In Table 1(a), we demonstrate that our poselet-based method is much more reliable than the state-of-the-art method [35]. An ablation analysis also reveals that all components in our human detection and tracking system jointly contribute to the superior performance.

[2] N depends on the aspect ratio of the poselet.

3.2 Learning saliency from gaze

Recall that it is important for our application to infer the salient people that the cameraman intends to capture, since there can be an unknown number of people (see Fig. 2 and 4(b)) in an unconstrained video. Given the tracklets, we aim to predict which one corresponds to a person performing salient actions. We train a predictor to generate a tracklet-based saliency score using multiple cues. To train our predictor, we asked the authors of the videos to watch them while recording eye gaze using a commodity gaze sensor (`http://www.mygaze.com`). We then used the gaze data as a measurement of ground truth saliency. We identify the person being salient in each frame when the gaze falls on a ground truth human annotation (see Fig. 3). We find that gaze tracks from the person who captured the video are much more informative as compared to a viewer who is unfamiliar with the event and people in it. Hence, we do not have such training data for videos from Youtube.

Our tracklet-based saliency score is built on top of a hypothesis-based saliency score. Here, we define a hypothesis-based saliency model which considers location, motion, and pose information of the hypotheses. Our training data consists of ground truth human annotations with binary "saliency" labels derived from gaze (at test time, we only have predicted human detections, and no gaze information). We train a random forest classifier to predict the saliency label and use the response of the classifier as the saliency score $s \in [0,\ 1]$ for each hypothesis using the following types of features.

Camera centric features. We define the camera centric features e as the location and relative height of the hypothesis with respect to the frame height. Hence, the model can learn the preferred location and scale of a salient person. For example, a salient person shouldn't be too small or too off-center. The feature also includes the Euclidean distance of the person's bounding box centroid to the frame center, which is typically used to model gaze in ego-centric videos [10,23].

Person motion features. We define the height changes $hr = h_t/h_{t+1}$ and motion direction (du, dv) in pixels between a pair of hypotheses (indices omitted) in two consecutive frames t and $t+1$ as the basic motion features $b = [hr, du, dv]$. This allows the classifiers to differentiate forward/backward and lateral motions, respectively. Our full motion features include motion uniqueness u derived from b and e (camera centric features) as follows (similar to the visual uniqueness measure in [34]),

$$u_i = \sum_j \|b_i - b_j\|^2 \cdot \omega(e_i, e_j) \ \ \textit{where} \ \ \omega(e_i, e_j) = \tfrac{1}{Z_i} \exp(-\tfrac{1}{2\sigma_p^2}\|e_i - e_j\|^2) \ , \quad (1)$$

where i, j are indices for hypotheses in the same frame, and Z_i is the normalization term. Hypothesis i has unique motion if its motion b_i is very different from the motion b_j of hypotheses at nearby location and scale (i.e., $e_i \sim e_j$).

Pose features. Pose provides a strong cue to determine an action. We use the raw poselet activation scores for each hypothesis as a coarse surrogate for pose.

Tracklet-level aggregation. The hypothesis-based saliency prediction is combined to produce the tracklet saliency score $\mathbf{s}_k$ (*s-score*) by summing up constituent scores $\{s_i\}_{i \in T_k}$, where T_k is a tracklet consisting of a set of hypotheses

and k is the tracklet index. In Table 1(b), we show that our tracklet-based saliency measure is more accurate in identifying salient people than a state-of-the-art video saliency estimator [38].

4 Salient Montages

Given the human tracklets with their saliency scores $\{(\mathbf{s}_k, T_k)\}_k$, we aim to generate salient montages ranked by their quality. In order to handle videos with various length, we first divide the tracklets into groups. There are multiple ways to generate groups. We use SIFT [27] point matching to find a group that is likely to contain tracklets appearing in physically nearby "places" (see technical report [40] for details).

Next, we introduce a unified model to (1) find a montageable moment in each group, (2) generate a montage for each group, and (3) rank the montages based on tracklet saliency and how visually pleasing they are. The overview of steps to generate salient montages is depicted in Fig. 2-Bottom.

4.1 Model

Our goal is to form a montage I_m from source images $\{I_i\}_{i\in\mathcal{L}}$, where the labeling space $\mathcal{L}$ is defined as the union set of hypotheses in all tracklets (i.e., $\cup_k\{i\}_{i\in T_k}$). Note that here we use the same index for both the source images and hypotheses for simplicity. This means that our model uses all hypotheses as candidate source images to generate a salient montage. More formally, we need to choose a source image index i, and a correspondence location $\hat{p}$ in the source image I_i for every pixel location p in the montage. Given i and $\hat{p}$, we assign the RGB value $(I_i(\hat{p}))$ of the source image to the RGB value $(I_m(p))$ of the montage. We define a pixel-level labeling variable L, where $i = L(p)$ denotes the source image index chosen at pixel location p in the montage. We also define a transformation M_i aligning the montage coordinate to the i source image coordinate such that $\hat{p} = M_i(p)$. The following montageability cost $C(L, \mathcal{M})$ (similar to [1]) is used to select the optimal pixel-level labeling L and transformations $\mathcal{M} = \{M_i\}_{i\in\mathcal{L}}$,

$$\min_{L,\mathcal{M}} C(L,\mathcal{M}) = \min_{L,\mathcal{M}} \textstyle\sum_p C_d(p, L(p); \mathcal{M}) + \sum_{p,q} C_I(p, q, L(p), L(q); \mathcal{M}) \ , \quad (2)$$

where C_d is the data term which encourages salient actions in source images to be selected, and C_I is the seam term considering color and gradient matching for reducing visual artifacts.

Instead of requiring users' annotations as in [1], we use the hypotheses locations and tracklet saliency to define the data term.

Saliency-based data term. Intuitively, we should include a hypothesis depicting a salient action in the montage. The cost of "not" including a pixel corresponding to hypothesis i in the montage depends on its saliency as follows,

$$C_d(p, \ell \neq i; \mathcal{M}) \propto \mathbf{s}_{k(i)} \cdot m_i(M_i(p)) \ , \quad (3)$$

where $\mathbf{s}_{k(i)}$ is the s-score of tracklet k containing hypothesis i, and $m_i(M_i(p))$ is the estimated probability that pixel $M_i(p)$ corresponds to hypothesis i (see

technical report [40] for more details). The final cost of pixel p assigned to hypothesis ℓ is defined as,

$$C_d(p, \ell; \mathcal{M}) = \lambda_d \max_{i \neq \ell} \mathbf{s}_{k(i)} \cdot m_i(M_i(p)) \ , \tag{4}$$

where we take the maximum cost of $i \neq \ell$, and λ_d is used to balance the seam term.

Seam term. Our seam term in color and gradient domains is defined as,

$$\begin{aligned} C_I(p, q, L(p), L(q); \mathcal{M}) = &||I_{L(p)}(\hat{p}) - I_{L(q)}(\hat{p})|| + ||I_{L(p)}(\hat{q}) - I_{L(q)}(\hat{q})|| \\ &+ ||\nabla I_{L(p)}(\hat{p}) - \nabla I_{L(q)}(\hat{p})|| + ||\nabla I_{L(p)}(\hat{q}) - \nabla I_{L(q)}(\hat{q})|| \ , \end{aligned} \tag{5}$$

where $\nabla I_i(p)$ is a 6-component color gradient (in RGB) of the source image i at pixel location p, both $\hat{p} = M_{L(p)}(p)$ and $\hat{q} = M_{L(q)}(q)$ are the transformed locations.

Before detailing how to solve Eq. 2 to obtain optimal labeling L^* and transformation $\mathcal{M}^*$, we discuss a way to rank montages from different groups of tracklets using a novel score derived from our model.

Montageability score. The minimum cost $C(L^*, \mathcal{M}^*)$ in Eq. 2 cannot be used to compare montages from different groups since the values are not normalized. To overcome this problem, we define a novel montageability score as,

$$V_{\mathcal{L}} = \frac{\min_{i \in \mathcal{L}} (C(L{=}i, \mathcal{M}^*))}{C(L^*, \mathcal{M}^*)} \tag{6}$$

where $L = i$ denotes $L(p) = i$ for all p which is a degenerate solution when only the i source image is used. The minimal degenerate solution is used to normalize the minimum cost so that this score is always larger than one. The larger the score, the better the quality of the montage with respect to the best degenerate solution. A very high score typically means many non-overlapping salient actions can be selected to reach a relatively low cost ($C(L^*, \mathcal{M}^*)$) compared to the best degenerate solution which only selects the most salient action.

Unique properties. Our model differs from [1] in two more ways: (1) both our data and seam terms depend on the transformation $\mathcal{M}$, and (2) our labeling space $\mathcal{L} = \cup_k \{i;\ i \in T_k\}$ is very large ($\sim 1K$) since it is the union of all hypotheses within each group. Due to these properties, it is challenging to jointly optimize transformations $\mathcal{M}$ and pixel-level labeling L. However, we observe that, when $\mathcal{M}$ is given and $\mathcal{L}$ is small, we can solve Eq. 2 using graph-cut efficiently. To this end, we first estimate $\mathcal{M}^*$ and obtain a small pruned set of hypotheses $\hat{\mathcal{L}}$ as the labeling space for computing the montage. A detailed overview of these steps is depicted in Fig. 2-Bottom. We describe each step in detail next.

4.2 Estimating transformations

We propose to search for transformations so that a maximum number of spatially non-overlapping hypotheses exist (implicitly reducing the cost of the data term). This is to ensure that many salient actions can be shown in the montage without blocking each other. We start by matching pairs of source images in

$F = \cup_k \{f(i)\}_{i \in T_k}$, where $f(i)$ denotes the source image index that hypothesis i appeared in. We introduce the f index here since many hypotheses appear in the same source image in practice. To efficiently match relevant pairs of images, we evenly divide all images into 1000 segments in time. All pairs within each segment and pairs across a few pairs of segments are matched (see technical report [40] for details). For each pair of images, we obtain sparse [27] and dense SIFT correspondences [24]. Given the correspondences between $\hat{f}$ and $f(i)$, we can estimate an initial 2D affine transformation $\hat{M}_{i,\hat{f}}$ to warp hypothesis i to frame $\hat{f}$ using correspondences surrounding the bounding box of the i hypothesis. The initial 2D affine transformation $\hat{M}_{i,\hat{f}}$ is then refined using Lucas-Kanade template matching [29][3]. Given the set of valid transformations $\hat{\mathcal{M}} = \{\hat{M}_{i,f}\}$, we calculate the binary connection matrix $Q = \{q_{i,f}\}$ such that $q_{i,f} = 1$ if $\hat{M}_{i,f}$ is valid. Next, we select the central image $f_c = \arg\max_f \sum_{i \in J_f} q_{i,f}$, where J_f is a set of mutually non-overlapping hypotheses at the f image coordinate (see technical report [40] for details). Finally, we obtain the transformation as $M_i^* \equiv \hat{M}_{i,f_c}$. Note that some of the hypotheses cannot be warped to the central image f_c due to limitations of existing registration methods. These hypotheses are removed from consideration in the next step.

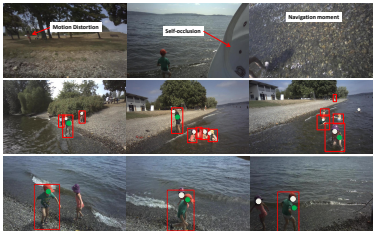


Fig. 3: Our data and annotations: the top row shows challenging frames from ego-centric videos: fast camera motion, wearer self-occlusion, and navigation moment. The middle row shows the ground truth person annotations (bounding boxes with person identity indices) and ground truth gaze (green dots). The bottom row shows ground truth salient people (red bounding boxes) and ground truth gaze (green dots). When bounding boxes overlap (bottom row, middle frame) we resolve ambiguity by minimizing identity switches.

Given: $R = \{r_{ij}\}$, and K pairs of tracklet and s-score $\{(T_k, \mathbf{s}_k)\}$ sorted from high to low saliency.
Return: a set of highly salient and mutually montageable hypotheses $\hat{\mathcal{L}}$ such that $r_{ij} = 1; \forall i, j \in \hat{\mathcal{L}}$
Initialize $\hat{\mathcal{L}}$ and $\mathcal{N}$ as empty.
Set tracklet index $k = 1$ to start with the most salient tracklet
repeat
  Select $i = \arg\max_{i \in T_k} r_i$, where $r_i = \sum_{j \notin \mathcal{N}} r_{ij}$.
  if $r_i \neq 0$ **then**
    Add i into $\hat{\mathcal{L}}$.
    Add $J = \{j;\ r_{ij} = 0\}$ into $\mathcal{N}$.
    Remove i and J from $\{T_k\}$.
  end if
  if T_k is empty or $r_i = 0$ **then**
    repeat
      k=k+1, which selects the next most salient tracklet.
    until T_k is not empty or $k > K$.
  end if
until $k > K$.

Algorithm 1: Greedy Hypotheses Selection.

[3] We remove pairs with transformations that have matrix condition larger than 1.25 to avoid bad registrations.

4.3 Selecting hypotheses

Recall that the number of hypotheses is typically too large for Eq. 2 to be solved efficiently. Hence, we need to select a small set of hypotheses $\hat{\mathcal{L}}$. We use the montageability score $V_{\{i,j\}}$ (defined in Eq. 6) of a pair of hypotheses (i, j) to decide if the pair is preferred to be jointly selected. If the montageability score $V_{\{i,j\}}$ is larger than β, we set $r_{ij} = 1$ which indicates that a pair of hypotheses (i, j) should be jointly selected. Given $R = \{r_{ij}\}$, the goal is to select a set of salient and mutually montageable hypotheses $\hat{\mathcal{L}}$ (i.e., $r_{ij} = 1$ for all $i, j \in \hat{\mathcal{L}}$) such that $\sum_{i \in \hat{\mathcal{L}}} \mathbf{s}_{k(i)}$ is maximized. In this way, we select as many salient hypotheses as possible by maintaining montageability. We greedily select a set of mutually montageable hypotheses $\hat{\mathcal{L}}$ using Algorithm 1. The algorithm proceeds down the list of tracklets sorted by s-score and adds those that are mutually montageable with currently selected ones. Note that we are implicitly reducing the cost of data term by maximizing the saliency of the selected hypotheses, and reducing the cost of the seam term by selecting mutually montageable set of hypotheses.

Once there are at least two selected hypotheses in $\hat{\mathcal{L}}$, we solve Eq. 2 to obtain L^* and generate the salient montage. We apply the same process for all groups, and use the montageability scores to retrieve salient montages with good quality.

5 Experiments

We evaluate our method on two types of videos. The first type includes unedited videos captured by two people using a head mounted camera (`www.looxcie.com`), where the camera center roughly follows the wearers' visual attention. The second type of videos are downloaded from Youtube. These are a mixture of edited and unedited videos captured mostly from hand-held devices. Both datasets are publicly available (see technical report [40] for details). We demonstrate that our method can handle both types of videos using the same settings.

In detail, we demonstrate that (1) our poselet-based people detection and tracking system outperforms the state-of-the-art method [35], (2) our saliency prediction method outperforms the state-of-the-art saliency detector [38], and (3) visually pleasing salient montages can be retrieved with high precision.

5.1 Family outing dataset

We collected a "family outing" dataset which contains 10 unedited videos with a total length of 2.25 hours ($243K$ frames). The videos include events in playgrounds, parks, lakeside, seaside, etc. These ego-centric videos are challenging for tracking and registration due to fast camera motion, self-occlusion from the wearer, and moments when the wearer is navigating the terrain (see Fig. 3-Top). We demonstrate that it is possible to generate impressive montages from these challenging videos. For training and evaluation, we collect ground truth human bounding boxes and gaze data from the camera wearers (see Fig. 3-Middle). We ask users on Amazon Mechanical Turk to annotate at least the three most salient people (if possible) for each frame using the video annotation system [42]. We also ask the camera wearers to watch the videos and record their gaze using a commodity gaze tracker. We use the gaze data from the camera wearer to assign

(a)	mAP	(b)	B1	B2	F	SS	CS	LS1	LS2	LS3	PG1	PG2	Avg.
[35]	5.09%	Our	38.51	**40.36**	**5.43**	**3.85**	**3.30**	**11.77**	**2.95**	**3.38**	**7.33**	**7.38**	**12.42**
Our1	9.28%	C. C. Feat.	45.06	44.39	16.73	18.39	4.75	13.09	5.18	5.92	14.69	11.24	17.95
Our2	17.70%	Raw Det. Score	139.27	195.77	22.57	13.75	7.53	37.21	13.78	10.45	16.03	11.19	46.75
Our3	**18.80%**	Video Saliency [38]	**33.93**	55.13	15.06	5.17	13.29	38.80	18.98	4.34	27.43	21.81	23.39

Table 1: (a) Tracking results: mean average precision (mAP) comparison of different human detection and tracking system. 'Our3" is our full poselet-based system. 'Our2" is our system without the poselet-based relative similarity. "Our1" is our system also without detection by poselet trajectories (i.e., linking person bounding boxes from poselet detectors.). (b) Ranking results: Weighted Spearman's footrule for rank error comparison. Our method (Our) achieves the smallest error except in "B1". On four videos, our average errors are lower than five. On average, our error is almost half of the [38]. C.C. denotes Camera Centric. B denotes beach. F denotes ferry. SS denotes seaside. CS denotes campsite. LS denotes lakeside. PG denotes Playground.

a ground truth binary "salient" label to each human annotation, assuming gaze reveals the salient subjects in the videos (see Fig. 3-Bottom).

Detecting and tracking humans. We optimize the parameters of our system on one video (Playground 1) and evaluate the performance on the remaining 9 videos. In Table 1 (a), we compare the detection accuracy using hypotheses in tracklets. Our full poselet-based method ("Our3") achieves a significant 13.71% improvement in mean average precision compared to [35] using the state-of-the-art human detector [11]. We also conduct an ablation analysis by removing components in our system one at a time (see Sec. 3.1 for the components). Our system without the poselet-based relative similarity ("Our2") becomes slightly worse than the full system. By further removing the detection by poselet trajectories, our system ("Our1") becomes much worse than the full system but is still better than [35].

Salient people prediction. It is critical for our method to discover the salient people in order to generate montages relevant to the camera wearer. Therefore, after we obtain tracklets, we need to rank the saliency of each tracklet. Given the ground truth salient label for each human annotation, we can generate the ground truth rank of the tracklets by counting the number of times the hypotheses in each tracklet overlapped[4] with the salient human annotation. Similarly, using the predicted s-scores $\mathbf{s}$ of each tracklet, we can also rank the tracklets. We conduct a leave-one-out cross-validation experiment by training the random forest classifier using nine videos and evaluating on the remaining one. Given the predicted rank and the ground truth rank, we calculate the weighted Spearman's footrule on the top 10 predicted tracklets to measure rank error as follows:

$$\frac{1}{W}\sum_{i=1}^{10} \|i-\rho(i)\| \cdot w_i \quad where \;\; w_i = \frac{1}{i},\; W = \sum_{i=1}^{10} w_i \;, \tag{7}$$

where i is the rank of the predicted tracklet, $\rho(i)$ is the ground truth rank of the i^{th} predicted tracklet, and w_i is the weight of each rank, which emphasizes the error of tracklets with higher predicted rank. We compare our prediction

[4] If the intersection area over the union area is bigger than 0.25, we consider two bounding boxes overlapping. If a salient human annotation is overlapping with more than one hypotheses, we consider only the tightest hypothesis.

(a) Playground2 (Youtube link) (b) Beach2 (Youtube link) (c) Lakeside1 (Youtube link)

Fig. 4: Montages from the family outing dataset. In each example, we show the montage and the selected frames overlaid with the hypotheses indicated by red bounding boxes.

using all the features with three baseline methods: (1) ranking using the raw sum of detection scores for each tracklet, (2) our method using only the camera centric features, and (3) ranking each tracklet using state-of-the-art video saliency detector [38]. For each frame, the video saliency detector generates a saliency probability map. We assign the median of the saliency probability within a bounding box as the saliency score for each hypothesis. Then we rank each tracklet using the sum of the saliency scores. A detailed rank error comparison is shown in Table 1 (b). Our method using all features significantly outperforms other baselines in 9 out of 10 videos, and our average rank error (12.42) is almost half of the error of [38].

Given the s-score of each tracklet, we generate montages as described in Sec. 4. We now discuss some interesting cases below.

Camera motion and scene depth variation. The montage A in Fig. 1 shows a little girl on a slide. Due to the lateral camera motion and the depth variation in the scene, it is extremely challenging to register these four images using a global transformation. Hence, typical methods [38,26] will fail. In contrast, our method can generate an impressive montage conveying the action. A similar result is shown in Fig. 4(a).

Crowded scenes. Videos often contain crowded scenes with multiple people. Fig. 4(b) shows that our method can select the most salient person in a crowd. A similar result is shown in Fig. 4(c).

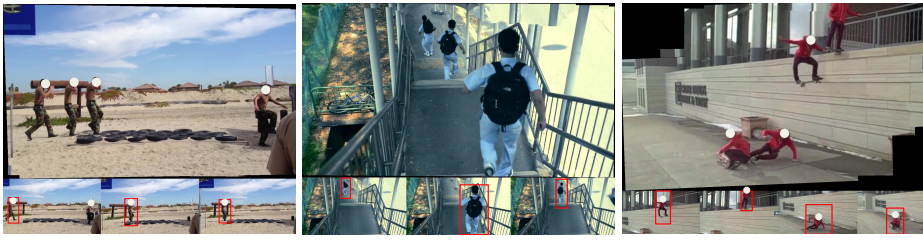
Our method generates on average about 18.1 montages per video in this dataset. Fig. 1 shows a set of montages our method retrieved to capture an afternoon at the playground. A lot more impressive salient montages and a video are included in the technical report [40].

5.2 Youtube dataset

We downloaded videos from Youtube following two procedures. First, we searched for less edited family outing videos using keywords such as “family outing”, “playground” and “park”. Then, we manually selected the first 6 less edited videos (an average of 10.4 minutes per video) and refer them as “Youtube outing dataset”. Second, we searched for three queries: “obstacle course”, “parkour”, and “skateboarding”, and downloaded the top three ranked creative commons licensed videos for each query. These nine videos (an average of 4.8 minutes per

(a) Youtube link (b)Youtube link (c) Youtube link

Fig. 5: Montages from the Youtube outing dataset. In panel (a), the most salient person is successfully selected. In panel (b,c), little kids are reliably detected and tracked.



(a) Youtube link (b) Youtube link (c) Youtube link

Fig. 6: Montages from the Youtube motion dataset. Our system nicely handles the fast motion in obstacle course, parkour, and skateboarding.

video) are referred to as "Youtube motion dataset". Note that videos in the motion data are typically edited.

Our method generalizes well to the Youtube outing dataset which is similar to the family outing dataset. Examples are shown in Fig. 5. We also achieve encouraging results on the Youtube motion dataset. Despite the fast motion and frame cuts, our system works reasonably well and generates fairly impressive results (Fig. 6). Please see technical report for more salient montages.

5.3 Retrieving good montages

In order to evaluate how well our method can automatically retrieve good montages, we ask human subjects to classify montages into *good*, *reasonable*, and *bad* ones. Good results are the ones without clear artifacts on the foreground objects and the actions are easy to understand, reasonable results are the ones with small artifacts on the foreground objects but the actions are still easy to understand, and bad results are the ones with either significant artifacts or the actions are not easy to understand. The raw distribution of good, reasonable, and bad montages are shown in Fig. 7. We evaluate how well our montageability score retrieves good montages. We achieve mean average precisions of 55%, 62%, and 78% compared to 44%, 58%, and 68% for a baseline method using only s-scores[5] on the family outing[6], Youtube outing, and Youtube motion[7] datasets. We also evaluate our recall rate on the challenging ego-centric family outing dataset. We ask a human subject to count the number of montageable moments

[5] For each montage, we sum the s-scores of the selected hypotheses for ranking.
[6] One video generates less than 10 montages and it was not included.
[7] Four videos generate less than 10 montages and they were not included.

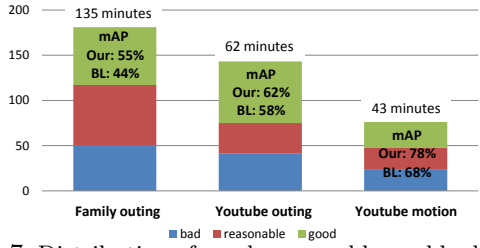


Fig. 7: Distribution of good, reasonable, and bad montages, where x axis is the number of montages. On top of each bar, we show the total length of videos in each dataset. The mean average precision (mAP) comparison between our method and the baseline (BL) method for retrieving good montages is overlaid on the bar plot.

Fig. 8: Failure cases. We show both the montage and the selected frames overlaid with the hypotheses indicated by red (correct ones) and blue (missing or incorrect ones) boxes. See technical report [40] for more analysis.

where there is a (not necessary salient) moving person not severely occluded by other objects or scene structures. Our system achieves an average recall of about 30%. Note that the low recall is fine for our application since, similar to a recommender system [7], we aim at showing users a few salient montages rather than all montageable moments which often might be boring.

5.4 Analysis of failure cases

Our automatic system inevitably returns some failure cases. Fig. 8-Left shows a typical failure case due to bad localization since the person appears in a rare pose. Fig. 8-Right shows a typical failure case where unselected people (depicted by a blue bounding box) in the scene are cut in half. A more robust tracking system and a montageability function incorporates information of unselected hypotheses can potentially resolve these cases. Finally, our system will not retrieve salient moments which are not montageable (e.g., severe camera translation).

Implementation details. Finally, we describe how we set the parameters in each component. We set all parameters for human detection and tracking on one video (Playground 1): minimal spatial overlap is set to $\rho = 0.25$, and the threshold for relative similarity is set to $\sigma = 1.5$. The bandwidth of motion uniqueness σ_p, the weight λ_d to balance the data and seam terms, and the threshold for montageability score β are empricially set to 5, 0.2, and 1.7, respectively. Our current Matlab implementation takes about 2 hours to process a 20 minutes video in family outing dataset on a single 8 cores machine.

6 Conclusion

We believe our method is the first to demonstrate the ability to automatically generate high quality montages from unconstrained videos. Our results on videos captured by wearable cameras are especially impressive due to the challenging conditions for tracking and registration methods to work reliably. In the future, we aim at inferring high-level semantic information in order to enable better prediction and understanding of "interesting moments".

Acknowledgement We thank Microsoft, Google, Intel, the TerraSwarm research center, NSF IIS-1218683, ONR N00014-13-1-0720, and ONR MURI N00014-10-1-0934 for supporting this research.

References

1. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. In: ACM SIGGRAPH (2004)
2. Aner, A., Kender, J.R.: Video summaries through mosaic-based shot and scene clustering. In: ECCV (2002)
3. Ballard, D.H.: Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition 13(2), 111–122 (1981)
4. Borgo, R., Chen, M., Daubney, B., Grundy, E., Heidemann, G., Hoferlin, B., Hoferlin, M., Janicke, H., Weiskopf, D., Xie, X.: A survey on video-based graphics and video visualization. In: EUROGRAPHICS (2011)
5. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
6. Chen, S., Fern, A., Todorovic, S.: Multi-object tracking via constrained sequential labeling. In: CVPR (2014)
7. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the Fourth ACM Conference on Recommender Systems. pp. 39–46. ACM, New York, NY, USA (2010)
8. Cui, X., Liu, Q., Metaxas, D.: Temporal spectral residual: fast motion saliency detection. In: ACM Multimedia (2009)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
10. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: ECCV (2012)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models, release 5. `http://www.cs.berkeley.edu/~rbg/latent/voc-release5.tgz`
12. Fragkiadaki, K., Zhang, W., Zhang, G., Shi, J.: Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In: ECCV (2012)
13. Goferman, S., Zelnik-Manor, L., Tal, A.: Contextaware saliency detection. TPAMI (2012)
14. Goldman, D., Curless, B., Salesin, D., Seitz, S.: Schematic storyboarding for video visualization and editing. In: SIGGRAPH (2006)
15. Gong, Y., Liu, X.: Video summarization using singular value decomposition. In: CVPR (2000)
16. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: CVPR (2008)
17. Irani, M., Anandan, P., Hsu, S.: Mosaic-based representations of video sequences and their applications. In: ICCV (1995)
18. Joshi, N., Metha, S., Drucker, S., Stollnitz, E., Hoppe, H., Uyttendaele, M., Cohen, M.F.: Cliplets: Juxtaposing still and dynamic imagery. In: UIST (2012)
19. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
20. Kalal, Z., Mikolajczyk, K., , Matas, J.: Tracking-learning-detection. TPAMI (2011)
21. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: CVPR (2013)
22. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)

23. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV (2013)
24. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: dense correspondence across difference scenes. In: ECCV (2008)
25. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. TPAMI (2010)
26. Liu, F., hen Hu, Y., Gleicher, M.: Discovering panoramas in web video. In: ACM Multimedia (2008)
27. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV (1999)
28. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR (2013)
29. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. Imaging Understanding Workshop (1981)
30. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. TPAMI (2010)
31. Massey, M., Bender, W.: Salient stills: Process and practice. IBM Systems Journal 35(3&4), 557–574 (1996)
32. Milan, A., Schindler, K., Roth, S.: Detection- and trajectory-level exclusion in multiple object tracking. In: CVPR (2013)
33. Ngo, C., Ma, Y., Zhan, H.: Video summarization and scene detection by graph modeling. In: CSVT (2005)
34. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)
35. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011)
36. Pritch, Y., Rav-Acha, A., Gutman, A., Peleg, S.: Webcam synopsis: Peeking around the world. In: ICCV (2007)
37. Rav-Acha, A., Pritch, Y., Peleg, S.: Making a long video short. In: CVPR (2006)
38. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: CVPR (2013)
39. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision (2009)
40. Sun, M., Farhadi, A., Seitz, S.: Technical report of salient montage from unconstrained videos. `http://homes.cs.washington.edu/~sunmin/projects/at-a-glace/`
41. Sunkavalli, K., Joshi, N., Kang, S.B., Cohen, M.F., Pfister, H.: Video snapshots: Creating high-quality images from video clips. IEEE Transactions on Visualization and Computer Graphics 18(11), 1868–1879 (2012)
42. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. IJCV pp. 1–21
43. Yang, B., Nevatia, R.: An online learned crf model for multi-target tracking. In: CVPR (2012)